

PROGRAMA DE CÁTEDRAS CONACYT 2015

DATOS GENERALES

Título / Nombre del proyecto: Análisis topológico de datos para matemáticas y sus aplicaciones

Institución: Centro de Investigación en Matemáticas, A.C.

Dependencia del Proyecto: CIMAT – Guanajuato

Entidad Federativa donde se realizará el proyecto: Guanajuato

Temática: Conocimiento del Universo

Reto: Estudios de física, matemáticas, química y sus aplicaciones

Modalidad: Grupal (3)

DESCRIPCIÓN DEL PROYECTO

Propósito:

Crear un núcleo de tres investigadores jóvenes de perfil multidisciplinario que contribuyan a formar un grupo destacado de investigación en Análisis Topológico de Datos (ATD), una temática científica emergente, innovadora, metodológicamente revolucionaria y que integra cuantiosos temas de matemáticas.

El ATD es en la actualidad una de las líneas de investigación más trascendentes en matemáticas, tanto por el conocimiento que genera como por el universo potencial de sus aplicaciones. Ofrece grandes retos no sólo para su campo de conocimiento, matemáticas básicas, sino también para otras áreas científicas del CIMAT: Computación y Estadística.

Su respaldo daría visibilidad internacional en el tema al CIMAT, a partir del impulso de los tres jóvenes especialistas y por el trabajo de los investigadores del CIMAT que se integren, sea por su interés en ATD o por su labor en otras técnicas de análisis de datos modernos; atendiendo una recomendación del Comité Evaluador Externo del CIMAT.

Objetivos:

Inducir la creación de un polo de investigación, intercambio académico y formación de recursos humanos, con habilidades y conocimientos multidisciplinarios para el análisis de datos complejos con diversos enfoques y novedosas combinaciones de ideas matemáticas, estadísticas y computacionales.

Desarrollar y utilizar métodos y modelos para colaborar en proyectos de investigación multidisciplinarios que requieren el reconocimiento de formas, superficies, agrupamientos, geometría, patrones y conexiones para obtener información útil en nubes de datos complejos surgidos en ciencias médicas y genómicas; redes de comunicación, sociales y económicas; servicios financieros y de negocios, ecología, energía, sector público, incluso dentro de la matemática misma, donde los objetos matemáticos pueden codificarse como nubes de datos para su uso universal, al describir propiedades estructurales e invariantes globales sin requerir modelos matemáticos explicativos explícitos.

Motivos de la institución para desarrollar el proyecto:

El ATD comenzó hace menos de dos décadas. Se fundamenta en la topología algebraica y computacional, planteando nuevos problemas de algoritmos, combinatoria, geometría, cómputo, análisis y gráficas, así como retos de inferencia estadística ad hoc. Contar con un grupo de investigación en ATD es de alta relevancia para el CIMAT, en vista del tamaño de las bases de datos modernas y de la proliferación de datos no estructurados, de alta dimensión y de gran volumen.

El CIMAT impulsó el ATD desde 2014 (ver <http://atd.cimat.mx/es>). Si bien en México no hay aún especialistas en ATD, el CIMAT tiene las condiciones idóneas para formar un grupo de investigación líder en el país que impulse integralmente estos estudios por contar con científicos altamente competentes para conjuntar líneas matemáticas que tradicionalmente se cultivan por separado. Se busca estimular el dinamismo en la comunicación interdisciplinaria, tendiente a educar y formar nuevos cuadros necesarios para su abordaje.

Contribución esperada de las cátedras:

La comisión de tres jóvenes investigadores especialistas en ATD es necesaria para el éxito de esta iniciativa, pues se constituirán como el núcleo del grupo de investigación, coadyuvarán a ampliar el espectro de problemas complejos que puede atender el CIMAT, en vista del tamaño de las bases de datos modernas, aprovechando la integralidad de las matemáticas básicas, computación y estadística en el CIMAT, así como su trayectoria en vinculación.

Su investigación estará dirigida hacia diversos aspectos de bases de datos complejas en cuanto a su volumen, calidad y estructura, en los cuales el análisis topológico de formas y superficies es determinante para extraer una información que genere el entendimiento de agrupamientos, patrones y relaciones entre variables.

Contribuir en el CIMAT a la diversificación de la matemática aplicada con un área emergente y de mucha actualidad que al mismo tiempo integra las tres áreas académicas del Centro.

Actividades:

Realizar investigación en nuevos métodos de topología algebraica para ATD, especialmente en aspectos multivariados y temporales de homología persistente.

Reunir investigadores en inferencia estadística y probabilidad para analizar y modelar aspectos de incertidumbre y variabilidad de resúmenes de persistencia y topología estocástica.

Un tema de relevancia teórica en ATD son los aspectos algebraicos, en particular los referentes a la teoría de representaciones de álgebras. En varios sitios en México, en particular en CIMAT, este tema se ha cultivado, y se cuenta ya con grupos interesados en la visita de expertos en ATD.

Convocar a investigadores a participar en un grupo de estudio en aspectos teóricos y algorítmicos de topología-geometría computacional.

Impartir cursos cortos de temas selectos de topología algebraica para estadísticos, probabilistas y computólogos, así como organizar cursos de temas cómputo matemático y estadística que son relevantes para el ATD.

Impartir cursos de temas de ATD que sean transversales a los posgrados del CIMAT.

Realizar reuniones de trabajo con colegas de instituciones que aborden problemas nacionales con retos en el análisis de bases de datos complejos, en particular en conservación de la biodiversidad, ciencias genómicas y de la salud, y políticas públicas.

Continuar promoviendo visitas de expertos en el mundo para fomentar colaboraciones con investigadores y estudiantes del país.

Instaurar un seminario periódico de ATD con participación de los jóvenes investigadores, investigadores del CIMAT y estudiantes de las tres áreas del centro.

Crear un seminario interinstitucional en el tema con participación y estudiantes de diversas instituciones de México.

Fomentar estancias posdoctorales en el CIMAT en temas generales de homología persistente y sus aspectos de probabilidad, estadística y cómputo matemático.

Organizar una Escuela anual en Análisis Topológico de Datos y Topología Estocástica.

Dirigir tesis de doctorado, maestría y licenciatura en estos temas, en codirección transversal con las tres áreas académicas del CIMAT y expertos nacionales y extranjeros.

Difundir los resultados de las investigaciones a través de congresos, eventos y la publicación de artículos.

Resultados e impactos esperados a 5 y 10 años:

Tanto al finalizar el primero como el segundo lustro, tener una actualización de líneas y proyectos de investigación acordes con los cambios tecnológicos y el estado de los datos vigentes. A partir de ello, plantear nuevos proyectos de investigación globales en los cuales, con el fin de enfrentarlos de manera cada vez más promisoriamente integral, se combinen las técnicas de ATD de manera teórica y práctica con otras técnicas emergentes para datos complejos, como lo son la estadística sobre variedades y el análisis de patrones basado en modelos deformables, entre otros.

En los primeros 5 años del proyecto, que el CIMAT sea el líder en la disseminación, la promoción y la investigación de ATD en el país y Latinoamérica, habiéndose incrementado el número de investigadores, alumnos, cursos de licenciatura y posgrado en programas de matemáticas y aplicaciones matemáticas estratégicas.

Graduar al menos a 6 estudiantes de licenciatura y maestría en los primeros cinco años y al menos a 12 en los siguientes cinco años. Asimismo, 5 estudiantes de doctorado en los primeros cinco años y al menos 8 en los siguientes cinco años.

Los graduados del doctorado, serán capaces de trabajar en problemas multidisciplinarios, con experiencia y conocimiento en las diversas áreas de las matemáticas que involucran el ATD: computación, estadística, topología algebraica y geometría.

Conseguir la codirección de tesis e impartición de cursos transversales a las tres áreas académicas del CIMAT.

Ser un referente en experiencias de trabajo multidisciplinario de frontera en otras ciencias, tecnología y el sector gubernamental, a través de la obtención de fondos de diversas convocatorias de fondos sectoriales y estatales del Conacyt, así como de cooperación internacional, para la contribución a resolver problemas nacionales.

Publicar un promedio de 1.5 artículos anuales por investigador en revistas de calidad internacional reconocida.

Tener por lo menos dos investigadores posdoctorantes por año, a partir del segundo, quienes se incorporen a otras instituciones nacionales una vez que concluyan sus estancias en el CIMAT.

Organizar en el CIMAT, o en sus unidades foráneas, una escuela anual de ATD para profesores e investigadores de instituciones de todo el país.

Durante los diez primeros años, organizar tres eventos internacionales especializados en ATD: al tercero, sexto y noveno año del proyecto.

Al concluir diez años, lograr el reconocimiento internacional del grupo de ATD del CIMAT, ser un referente para la investigación en problemas vanguardistas en ATD, abordando nuevos retos en el manejo, utilización, análisis e interpretación de bases de datos.

Vinculación, pertinencia y perspectivas de transferencia de tecnología:

En los últimos 15 años, en el extranjero han surgido empresas que ofertan como servicio la interpretación de bases de datos usando la metodología del ATD. Estas compañías han sido muy exitosas y fueron fundadas por topólogos algebraicos y computacionales. Una de ellas, Ayasdi (www.ayasdi.com), es en este 2015 una de las 6 empresas más innovadoras en Big Data. Recientemente, egresados de doctorado especializados en ATD se han insertado fácilmente en industrias y compañías de análisis de datos.

Puede verse un ejemplo sencillo de cómo los métodos topológicos se usan para estudiar bases de datos complejas de alta dimensionalidad y volumen en el artículo “Extracting insights from the shape of complex data using topology” de Gunnar Carlsson et. al., <http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html>

CIMAT es una de las instituciones de matemáticas en México y Latinoamérica con más experiencias exitosas en la vinculación con los sectores académico, público y privado. El presente proyecto incrementaría la oferta y la capacidad del CIMAT con métodos emergentes a nivel mundial de matemáticas, estadística y computación, relevantes para el análisis y la extracción de información de bases de datos complejas.

En particular, se cuenta con una iniciativa para explorar el uso de ATD en el llamado problema de predicción del nicho ecológico. Consiste de explorar la estructura (topológica) del espacio ambiental para explicar la presencia de una especie. Tiene implicaciones en la predicción de especies invasivas, el estudio de especies patógenas, la conservación de especies; pertinente dentro del Problema Nacional de Aprovechamiento y protección de ecosistemas y de la biodiversidad del PECiTI del Conacyt.

Otro proyecto en el que se ha trabajado es en la estructura de la red de citas de artículos científicos. Se cuenta con gran cantidad de información estadística (de disciplinas, grupos de investigación, países, etc.) que requiere análisis tanto estadístico como en aspectos teóricos (teoría espectral de gráficas y redes complejas). Esto tiene el potencial de ser aplicado a estudios de políticas de evaluación de impacto de la investigación en México.

Las neurociencias, en especial el estudio de la topología de zonas de activación y patrones de conectividad a partir de neuroimágenes, son otra esfera en la cual este proyecto es pertinente. Como antecedente, existe una amplia trayectoria de colaboración entre el CIMAT y el Instituto de Neurobiología de la UNAM, en investigaciones multidisciplinarias acerca de nuevos métodos de análisis de neuroimágenes.

Además, como parte de la constante colaboración con el INEGI, en junio de 2014 el CIMAT organizó tres conferencias sobre ATD y Minería de Datos dentro del “Seminario Internacional: Big Data para la información oficial y la toma de decisiones”. Varios de las exposiciones de otros ponentes de este seminario hacen prever un gran número de posibles aplicaciones de ATD en México.

Descripción del grupo de investigación asociado al proyecto:

Participan investigadores de las tres áreas académicas del CIMAT: Matemáticas Básicas (MB), Probabilidad y Estadística (PE), y Ciencias de la Computación (CC), además de académicos de la Universidad de Guanajuato (UG).

Los investigadores solicitados constituirían un nodo central alrededor del cual se desprenderían dos ejes principales: la matemática básica y el complemento algorítmico-computacional y probabilístico-estadístico. El rol de los otros integrantes sería diverso, desde el interés en aspectos metodológicos de ATD (*) a la contribución con otras técnicas de análisis de datos y sus aplicaciones (+).

Dr. Octavio Arizmendi, PE, SNI-1, (*), gráficas, probabilidad.

Dr. Rolando Biscay, PE, SNI nuevo ingreso, (*), estadística, multidisciplinaria.

Dr. Gonzalo Contreras, MB, SNI-3, sistemas dinámicos.

Dr. José Antonio de la Peña, MB, SNI-3, (*), álgebra, redes.

Dra. Eloísa Díaz Francés, PE, SNI-2, (+), estadística, multidisciplinaria.

Dra. Claudia Esteves, UG, SNI-1, (+), geometría computacional.

Dr. José Carlos Gómez Larrañaga, MB, SNI-3, (*), topología geométrica y algebraica.

Dr. Rogelio Hasimoto, CC, SNI-1, (+), comunicaciones internet, reconocimiento de objetos 3D.

Dr. Jean Bernard Hayet, CC, SNI-1, (+), visión computacional.

Dr. Miguel Nakamura, PE, SNI-2, (*), estadística, multidisciplinaria.

Dr. Víctor Núñez, MB, SNI-1, (*), topología.

Dr. Víctor Pérez Abreu, PE, SNI-3, (*), probabilidad.

Dr. Enrique Ramírez, MB, SNI-1, (*), topología.

Dr. Johan Van Horebeek, CC, SNI-2, (+), reconocimiento estadístico de patrones, aprendizaje máquina.

Dr. Enrique Villa D., PE, SNI-1, (+), estadística, vinculación con la industria.

Dr. Ramón Reyes C., Centro CONACYT INFOTEC, (+), aspectos computacionales.

Asimismo, los 5 ponentes expertos que participaron en la Escuela de ATD de enero 2015 (<http://atd2015.eventos.cimat.mx/>) expresaron su deseo de realizar estancias en el CIMAT en los próximos años.

Descripción de la infraestructura física disponible para ejecutar las actividades del proyecto:

La sede principal de CIMAT, en Guanajuato, está asentada en una superficie 19 mil 420 m², dentro de los cuales se localizan 8 laboratorios, 1 auditorio y un salón de usos múltiples con capacidad de videoconferencia, 7 salas de juntas, 2 salas de videoconferencias, 13 salones de seminarios, 6 salones de usos múltiples, 5 centrales de datos y 150 cubículos con equipo de cómputo.

La biblioteca del Centro concentra un acervo de más de 28 mil 836 libros y más de 700 tesis, una colección de 674 revistas científicas: 234 suscripciones vigentes, 123 con acceso electrónico y 33 bibliotecas digitales y bases de datos.

A través de cuatro enlaces externos, se tiene acceso a la red de internet normal y a internet 2, que nos mantiene intercomunicados con otras instituciones académicas de México y el extranjero. Se cuenta también con servicios de internet comercial para regular el tráfico del servicio principal de red, y para el servicio seguro y eficaz de internet inalámbrico a toda la institución.

Por lo que se refiere a telefonía, el sistema conmutador de nuestra sede en Guanajuato cuenta actualmente con 200 extensiones analógicas y digitales y 180 extensiones IP en uso y tiene capacidad de crecimiento hasta un mil 150 extensiones; dispone, además, de correo de voz y operador automático, conferencia tripartita, grupos de telefonía y marcación directa. El sistema de conmutador cuenta con interconexión directa a los conmutadores de las Unidades Aguascalientes, Zacatecas y próximamente Monterrey.

El CIMAT tiene un clúster de HPC que permite hacer cómputo de alto desempeño, procesar gran cantidad de información en poco tiempo y resolver problemas de gran envergadura a través de un mil 190 C (230 cores + 960 cores (sin HT)) y 3.6TB RAM, complementado con máquinas con GPUs.

Se dispone además de un centro de hospedaje (Cimatel) para 50 personas distribuidas en 29 habitaciones, todas ellas con acceso a internet inalámbrico. Esta infraestructura se utilizará para la organización de las escuelas y talleres que se plantean en este proyecto.

Relación del proyecto con algún laboratorio nacional:

Recientemente, uno de los investigadores del CIMAT participantes en el proyecto realizó una estancia sábrica en el Centro de Investigación y Docencia en Economía (CIDE). Ello potenció la colaboración con el Laboratorio Nacional de Políticas Públicas de ese centro, mismo que cuenta con un banco de información que entre sus objetivos está el de apoyar la investigación que genere análisis y reportes útiles, tanto para la academia como para otros sectores.

El CIDE tiene, además, otros proyectos ambiciosos que exigen el manejo de grandes bases de datos, pues sin un manejo adecuado de ellas actualmente no se entiende la investigación en ciencias sociales y la elaboración de políticas públicas. Por ejemplo, la mencionada compañía AYASDI realiza estudios acerca del impacto de los programas de salud pública en algunos países en desarrollo. Nuestra iniciativa tiene el potencial de formar recursos humanos y grupos de colaboración capaces de realizar un trabajo similar.

Relación del proyecto con los programas registrados en el PNP:

El CIMAT ofrece programas de maestría y doctorado en Ciencias con orientación en las tres vertientes del proyecto: Matemáticas Básicas y Aplicadas, Probabilidad y Estadística y Ciencias de la Computación.

Todos estos posgrados tienen nivel de competencia internacional en el Padrón Nacional de Posgrados de Calidad de Conacyt. Estos programas atraen cada vez más alumnos de todo el país y del extranjero. Actualmente se tiene una matrícula de 150 alumnos de maestría y 82 de doctorado, siendo 41 extranjeros en todos los posgrados.

En particular, la maestría en probabilidad y estadística ha aumentado en las últimas cuatro promociones de 12-14 admitidos a 20-22, manteniendo un índice de eficiencia terminal alta; el doctorado ha incrementado en 30% su ingreso.

Al contarse con un grupo de investigación integral en ATD se abrirá de manera transversal una nueva línea en esos posgrados, en un tema con un creciente número de aplicaciones, en una diversidad de áreas y problemas de relevancia nacional.

Los posgrados del CIMAT tienen una fuente natural de estudiantes provenientes de la Licenciatura en Matemáticas de la Universidad de Guanajuato, la cual se ofrece en convenio con el CIMAT y en donde los alumnos reciben una formación integral. Históricamente esta licenciatura es un insumo del posgrado y, en ese sentido, hay en progreso algunas tesis de licenciatura enfocadas en el tema de ATD que tienen el potencial para convertirse temáticamente en tesis de maestría y doctorado.

Por otra parte, se fomentará el trabajo transversal del grupo de ATD, beneficiando al Centro con una formación actual de los maestros y los doctores.

Algunos investigadores en líneas nuevas, pero con vínculos fuertes y complementarios con los grupos existentes, impartirán cursos de posgrados en temas de análisis de datos complejos. Con ello se fomentará la creación de nuevos grupos de investigación en el tema y la formación de recursos humanos especializados en el análisis de datos complejos que puedan integrarse al mercado laboral, generando un círculo virtuoso para la institución y para el país.

Un proyecto como el que se propone puede dar como resultado la formación de recursos humanos que hagan posible que empresas en México ofrezcan esta nueva tecnología para la utilización de datos.

Es relevante mencionar que la Escuela de Análisis Topológico de Datos y Topología Estocástica organizada por el CIMAT en enero de este año tuvo una asistencia récord de 110 participantes. Esto incrementó el interés en el tema por

parte de un gran número de alumnos nacionales y extranjeros; algunos de ellos manifestaron su deseo de estudiar un posgrado que aborde estas temáticas novedosas y transversales de la matemática.



PERFILES SOLICITADOS

1. Grado académico: Doctorado.

Área: 2.1 Ciencias Físico-Matemáticas y Ciencias de la Tierra.

Disciplina: Matemáticas.

Subdisciplina: Topología.

Experiencia profesional:

Doctorado con no más de ocho años de haber sido obtenido.

Trabajo de investigación de preferencia en topología algebraica y con interés en aprender ATD.

Producción de artículos en revistas de reconocido prestigio internacional de acuerdo con la antigüedad de su doctorado.

Experiencia docente e interés y compromiso en perfeccionar su calidad didáctica para exponer a sus colegas y alumnos temas que mezclen áreas diferentes de ATD.

Perfil de nivel 1 del Sistema Nacional de Investigadores.

Preferentemente, con experiencia en la colaboración con científicos de otras disciplinas que poseen datos complejos para su análisis.

Actividades a desarrollar:

Participar en el grupo de trabajo del proyecto.

Realizar investigación en análisis topológico de datos, en particular acerca de la estabilidad algébrica de módulos de persistencia y aplicaciones de topología algebraica, tales como la teoría de obstrucción, para obtener información cualitativa de bases de datos de alto volumen.

Colaborar con los otros integrantes del proyecto, en particular con los de topología, en relación con nuevos modelos y métodos topológicos para el análisis de datos complejos. En especial, dado que las colecciones de objetos matemáticos se pueden codificar como nubes de datos, determinar si el ATD devela propiedades matemáticas significativas de los objetos de estudio.

Participar en el seminario periódico de ATD del CIMAT.

Impartir un curso semestral de topología algebraica aplicada.

A partir del segundo año, dirigir una tesis de posgrado en temas de topología algebraica aplicada.

Trabajar a partir del segundo año en proyectos multidisciplinarios.

2. Grado académico: Doctorado

Área: Ciencias Físico-Matemáticas y Ciencias de la Tierra

Disciplina: Matemáticas

Subdisciplina: Estadística.

Experiencia profesional:

Doctorado con no más de ocho años de haber sido obtenido.

Trabajo de investigación de doctorado o posdoctorado, de preferencia en inferencia estadística para datos complejos o geometría y topología estocástica.

Producción de artículos en revistas de reconocido prestigio internacional de acuerdo con la antigüedad de su doctorado.

Experiencia docente e interés y compromiso en perfeccionar su calidad didáctica para exponer a sus colegas y alumnos temas que mezclen áreas diferentes de ATD.

Perfil de nivel 1 del Sistema Nacional de Investigadores. Contar con conferencias invitadas.

Preferentemente, con experiencia en la colaboración con científicos de otras disciplinas que tienen datos complejos para su análisis.

Actividades a desarrollar:

Participar en el grupo de trabajo del proyecto.

Realizar investigación en inferencia estadística y probabilidad para el análisis topológico de datos, en particular en relación con inferencia estadística para superficies y variedades y selección de modelos de representación de datos

Colaborar con los otros investigadores del proyecto, en particular con los del área de probabilidad y estadística, en aspectos de incertidumbre y variabilidad del análisis topológico de datos, especialmente en inferencia estadística para superficies y formas, y a partir de resúmenes de homología persistente como diagramas, panoramas y códigos de barras.

Participar en el seminario periódico de ATD del CIMAT.

Impartir un curso semestral en el tema de inferencia estadística del análisis topológico de datos.

A partir del segundo año, dirigir una tesis de posgrado en temas de estadística del análisis topológico de datos.

Trabajar, a partir del segundo año, en proyectos multidisciplinarios.

3. Grado académico: Doctorado

Área: Ciencias Físico-Matemáticas y Ciencias de la Tierra

Disciplina: Matemáticas

Subdisciplina: Otras subdisciplinas de Matemáticas

Experiencia profesional:

Doctorado con no más de ocho años de haber sido obtenido.

Trabajo de investigación de doctorado o posdoctorado, de preferencia en topología algebraica aplicada o computación.

Tener sólidas bases de topología algebraica, experiencias significativas en ciencias de la computación e interés en aprender ATD y aprendizaje de máquina.

De preferencia, con experiencia posdoctoral o producción científica equivalente.

Producción de artículos en revistas de reconocido prestigio internacional, de acuerdo con la antigüedad de su doctorado.

Experiencia docente e interés y compromiso en perfeccionar su calidad didáctica para exponer a sus colegas y alumnos temas que mezclen áreas diferentes de ATD.

Perfil de nivel 1 del Sistema Nacional de Investigadores. Contar con conferencias invitadas.

Preferentemente, con experiencia en la colaboración con científicos de otras disciplinas que tienen datos complejos para analizarlos.

Actividades a desarrollar:

Participar en el grupo de trabajo del proyecto.

Realizar investigación en aspectos teóricos y algorítmicos de topología computacional, en particular en el desarrollo de algoritmos eficientes de cálculo y visualización de homología persistente multidimensional.

Colaborar con los otros investigadores del proyecto, en particular con los del área de computación, en aplicaciones computacionales de ATD y aspectos computacionales de ciencias de redes y topología.

Participar en el seminario periódico de ATD del CIMAT.

Impartir un curso semestral en el tema de topología computacional.

A partir del segundo año, dirigir una tesis de posgrado en temas de topología computacional.

Trabajar, a partir del segundo año, en proyectos multidisciplinarios.